1. What is the definition of a target function? In the sense of a real-life example, express the target function. How is a target function’s fitness assessed?

In the context of machine learning and optimization algorithms, a target function, also known as an objective function or fitness function, is a mathematical function that defines the goal or desired outcome of a problem. It represents the quantity that needs to be maximised or minimised in order to achieve the desired result. A target function can vary depending on the specific problem and its objectives. It could be a function that measures the accuracy of a predictive model, the error or cost to be minimised, the profit to be maximised, or any other metric that reflects the desired outcome.

2. What are predictive models, and how do they work? What are descriptive types, and how do you use them? Examples of both types of models should be provided. Distinguish between these two forms of models.

**Predictive Models:**

Predictive models aim to forecast or predict future outcomes based on historical data and patterns. These models utilise various algorithms and statistical techniques to identify relationships between input variables (features) and the target variable (the variable to be predicted). The predictive models learn from historical data to make predictions on new, unseen data.

**Descriptive Models:**

Descriptive models, on the other hand, aim to summarise and understand existing data patterns and relationships. These models focus on describing the data and extracting insights from it, rather than making future predictions. Descriptive models help in identifying trends, patterns, and correlations in the data, enabling better understanding of the underlying phenomenon.

Distinguishing between Predictive and Descriptive Models:

Goal: Predictive models focus on making predictions about future outcomes, whereas descriptive models aim to summarize and understand existing data patterns.

Time Orientation: Predictive models utilize historical data to make predictions about future events, while descriptive models analyze past data to extract insights about the observed phenomena.

Usage: Predictive models are used for forecasting, risk assessment, decision-making, and personalized recommendations. Descriptive models are used for exploratory data analysis, understanding relationships, and gaining insights.

Outputs: Predictive models produce predictions or probability estimates, while descriptive models generate summaries, visualizations, or classifications based on existing data patterns.

3. Describe the method of assessing a classification model’s efficiency in detail. Describe the various measurement parameters.

Accuracy:

Accuracy measures the overall correctness of the model's predictions. It calculates the ratio of correct predictions to the total number of predictions. While accuracy is a commonly used metric, it may not be appropriate for imbalanced datasets where one class dominates the others.

Precision:

Precision measures the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive (true positives + false positives). It focuses on the quality of positive predictions and is particularly useful when the cost of false positives is high.

Recall (Sensitivity or True Positive Rate):

Recall measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances (true positives + false negatives). It assesses the model's ability to identify all positive instances and is useful when the cost of false negatives is high.

F1 Score:

The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of both precision and recall and is useful when you want to consider both metrics equally. It is calculated as: F1 = 2 \* (precision \* recall) / (precision + recall).

Specificity (True Negative Rate):

Specificity measures the proportion of correctly predicted negative instances (true negatives) out of all actual negative instances (true negatives + false positives). It is particularly important when the cost of false positives is high.

Area Under the ROC Curve (AUC-ROC):

The ROC (Receiver Operating Characteristic) curve plots the true positive rate (recall) against the false positive rate. The AUC-ROC represents the area under this curve and provides a measure of the model's ability to distinguish between classes. Higher AUC values indicate better performance.

Confusion Matrix:

A confusion matrix is a tabular representation of the model's predictions against the actual class labels. It shows the counts of true positives, true negatives, false positives, and false negatives. It provides a more detailed view of the model's performance and can be used to calculate various metrics mentioned above.

4.

i. In the sense of machine learning models, what is underfitting? What is the most common reason for underfitting?

In the context of machine learning models, underfitting refers to a situation where a model is not able to capture the underlying patterns and relationships present in the training data. An underfit model performs poorly not only on the training data but also on new, unseen data.

The most common reason for underfitting is a model that is too simple or lacks complexity. This typically occurs when the model is not able to represent the underlying complexity of the data. Some common causes of underfitting include:

1. Insufficient Model Complexity: If the model is too simple, it may not have enough capacity to capture the intricate relationships in the data. For example, using a linear regression model to fit a highly non-linear relationship may result in underfitting.

2. Insufficient Features: If the model lacks important features or variables that are necessary for capturing the patterns in the data, it may lead to underfitting. Inadequate feature selection or feature engineering can contribute to this issue.

3. Over-regularization: Excessive regularization, such as using a high regularization parameter in techniques like Ridge Regression or Lasso Regression, can overly constrain the model and result in underfitting. Regularization is useful for preventing overfitting, but too much regularization can lead to a model that is too rigid and unable to capture the data's complexity.

4. Limited Training Data: When the training dataset is small or unrepresentative of the underlying population, the model may struggle to learn the true patterns. Insufficient training data can limit the model's ability to generalise well to new, unseen data, leading to underfitting.

To address underfitting, some potential solutions include:

- Using more complex models that have higher capacity to capture complex relationships in the data, such as deep learning models or ensemble methods.

- Increasing the number of relevant features or performing feature engineering to provide more information to the model.

- Reducing regularisation or adjusting hyperparameters to strike a better balance between model complexity and generalisation.

- Collecting more diverse and representative training data to improve the model's ability to learn the underlying patterns.

It's important to note that finding the right balance between model complexity and generalization is crucial to avoid both underfitting and overfitting, ultimately leading to better model performance and generalization to new data.

ii. What does it mean to overfit? When is it going to happen?

Overfitting occurs when a machine learning model learns the training data too well to the point that it starts to memorize noise and irrelevant patterns specific to the training set. As a result, the overfit model performs exceptionally well on the training data but fails to generalize to new, unseen data.

Overfitting tends to happen in the following scenarios:

1. Insufficient Training Data: When the available training data is limited, the model may overfit by trying to fit noise or outliers present in the small dataset. The model essentially "memorizes" the training examples without capturing the underlying patterns.

2. Overly Complex Model: If the model is excessively complex with too many parameters, it can lead to overfitting. A highly flexible model can easily capture noise and irrelevant details in the training data, resulting in poor generalization.

3. Feature Overload: When the model is trained with a large number of irrelevant or redundant features, it becomes more susceptible to overfitting. Including too many irrelevant features can mislead the model and cause it to learn spurious correlations.

4. Lack of Regularization: Regularization techniques, such as L1 or L2 regularization, are used to prevent overfitting by adding a penalty term to the model's objective function. If no or insufficient regularization is applied, the model may overfit by not properly controlling the complexity of the learned patterns.

5. Data Leakage: Data leakage occurs when information from the test or validation set unintentionally influences the training process. If the model inadvertently learns from this leaked information, it can lead to overfitting.

To address overfitting, several techniques can be applied:

1. Increase Training Data: Gathering more diverse and representative training data can help the model to generalize better and reduce overfitting.

2. Feature Selection: Selecting relevant features and removing redundant or irrelevant ones can improve the model's performance and reduce overfitting.

3. Regularization: Applying regularization techniques, such as L1 or L2 regularization, can help prevent overfitting by adding a penalty term to the model's objective function, encouraging simplicity.

4. Cross-Validation: Utilizing cross-validation techniques, such as k-fold cross-validation, can provide a more reliable estimate of the model's performance by evaluating it on multiple subsets of the data.

5. Early Stopping: Monitoring the model's performance during training and stopping the training process when the model starts to overfit can help prevent further deterioration of generalization performance.

By addressing overfitting, models can achieve better generalization and perform well not only on the training data but also on unseen data, which is the ultimate goal of machine learning.

iii. In the sense of model fitting, explain the bias-variance trade-off.

The bias-variance trade-off can be visualized as follows:

High Bias, Low Variance: Models with high bias but low variance tend to be overly simplified and make strong assumptions about the data. They may consistently underperform but are less prone to overfitting. Examples include linear regression with few features or simple decision trees.

Low Bias, High Variance: Models with low bias but high variance are highly flexible and capable of capturing complex relationships in the data. However, they are more susceptible to overfitting due to their sensitivity to noise and fluctuations in the training data. Examples include deep neural networks or highly complex ensemble models.

5. Is it possible to boost the efficiency of a learning model? If so, please clarify how.

6. How would you rate an unsupervised learning model’s success? What are the most common success indicators for an unsupervised learning model?

7. Is it possible to use a classification model for numerical data or a regression model for categorical data with a classification model? Explain your answer.

8. Describe the predictive modelling method for numerical values. What distinguishes it from categorical predictive modelling?

9. The following data were collected when using a classification model to predict the malignancy of a group of patients’ tumours:

i. Accurate estimates – 15 cancerous, 75 benign

ii. Wrong predictions – 3 cancerous, 7 benign

Determine the model’s error rate, Kappa value, sensitivity, precision, and F-measure.

10. Make quick notes on:

1. The process of holding out

2. Cross-validation by tenfold

3. Adjusting the parameters

11. Define the following terms:

1. Purity vs. Silhouette width

2. Boosting vs. Bagging

3. The eager learner vs. the lazy learner